MLP 군집 모델에 기반한 어구독립 화자증명

이태승° 최호진
한국항공대학교 {항공전자공학과, 전자정보통신컴퓨터공학부}
thestaff@hitel.net hjchoi@mail.hangkong.ac.kr

Text-Independent Speaker Verification Based on MLP Cohort Model

Tae-Seung Lee° Ho-Jin Choi
{Dept. of Avionics, School of Electronics Telecommunication Computer Engineering}, Hankuk Aviation University

요 약

본 논문에서는 기존의 확률적 화자군집 모델을 MLP(multi-layer perceptron)로 구현하는 방법과 원형 화자군집 모델이 갖는 문제를 해결할 수정 모델을 제시한다. 화자군집 모델은 화자등록 시간에 민감한 실용 환경에서 중요한 의미를 지닌다. 본 연구에서 사용한 인식단위는 여러 음소계열에서 지속적인 부분을 추출한 지속음이므로 화자등록과 증명 단계에서 특정한 어구에 한정되지 않는 어구독립 방식을 채택한다.

1. 서론

생체 인식에 대한 관심이 고조되는 요즘 음성에 의한 화자 선별 및 증명(identification & verification)에 대한 연구가 활발히 진행되고 있다. 본 논문에서는 최근 주목받고 있는 확률적 화자군집 평준화 방식(speaker cohort normalization)을 MLP(multi-layer perceptron) 영역에서 구현하려는 시도를 서술한다. 본 연구에서 채택한 화자증명 방식은 단모음을 포함하는 지속음을 판별 대상으로 삼기 때문에 화자등록과 증명시도의 단계에서 특정한 단어(word) 또는 어구(text)에 구애되지 않는 어구독립 증명이 가능하다.

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서 화자증명에 이용되는 지속음을 설명하고, 3장에서 화자군집 평준화 방식을 소개하며, 4장에서 확률적 화자군집 모델을 MLP로 구현하는 방법을 제안한다. 5장에서는 음성 데이터베이스를 이용하여 세 가지 방법으로 한정된 영역에서 수행한 실험 결과를 제시하고, 6장에서 본 연구의 결과를 정리한다.

2. 지속음 대상 화자증명

화자증명은 어구 요구의 방식에 따라 어구 종속/독립/요구(text-dependent/independent/prompt)로 분류된다. 어구 종속은 화자 등록과 인식시 동일한 어구 사용을 강제하고, 어구 독립은 등록과 증명마다 자유로운 어구를 사용하며, 어구 요구는 등록에는 자유로운 어구를 사용하지만 증명시 특정한 어구를 인식기에서 요구하는 방식이다. 보안면에서 어구 종속과 독립 방식은 녹음된 데이터를 통해 사칭될 우려가 크다. 그러나 어구독립을 음소 단위 인식으로 구현한다면 음성인식 능력을 추가하여 어구요구 방식으로 개량할 수 있다.

화자인식에서는 음소마다 화자간 차이를 구별하는 능력에 차이가 있다. 지금까지의 연구 결과[Eatock94][Delacretaz98]에 따르면 비음, 모음, 마찰음, 폐쇄음 순서로 화자 인식에 기여하는 것으로 나타났다. 이들 연구에서는 영어의 기본 음소를 대상으로 인식률을 검사했지만, 비음>모음>마찰음>폐쇄음의 순위를 감안할 때나 일상적인 청각인지 경험을 고려했을 때 우리의 청각 능력에서 화자인식이 음성의 지속적인 부분의 화자간 차이를 바탕으로 한다고 가정할 수 있다. 이에 따라, 본 연구에서는 지속적인 부분이 비교적 많은 비음, 모음, 마찰음에서 지속부분을 채취하여 화자간 인식의 기본 단위로 사용한다.

3. 화자군집 평준화 방식[Furui96]

음성인식과 달리 화자인식에서 인식점수는 비교 대상이 되는 화자 집단이 명확히 정의되어 있지 않기 때문에 화자 사이의 단순 비교가 어렵다. 따라서 어떤 방식으로든 의뢰화자(client)를 다른 화자로 평준화시킨 후 그 값에 일정한 문턱값(threshold)을 적용하여 최종적인 판단을 내리게 된다. 이 때 평준화에 사용되는 화자 모델에 따라 광역 모델(world model)과 지역 모델(cohort model)로 나누거나 이 둘을 혼합하여 적용한다[Gu99].

평준화 방식은 다음과 같이 유사도 비율(likelihood ratio)의 수식으로 표현된다.

$$\log L(X) = \log p(X|S=S_c) - \log p(X|S\neq S_c) \qquad (1)$$

여기서 S는 배경화자이고 S_c는 주장화자이다. 일반적으로 $\log L$의 값이 양수이면 증명된 것으로, 음수이면 사칭한 것으로 판단한다.

배경화자 집단이 모든 화자를 충분히 대표하고 있다고 가정할 때, 주장화자 S_c외의 다른 모든 화자에 대해 X 지점의 확률 밀도는 최근접 참조화자의 확률 밀도에 의해 좌우될 가능성이 있다. 따라서 식 (1)을 다음과 같이 변경할 수 있다.

$$\log L(X) = \log p(X|S=S_c) - \max_{S \in Ref, S \neq S_c} \log p(X|S \neq S_c) \quad (2)$$

그러나, 이 결정식은 두 가지 이유 때문에 현실적이지 못하다. 즉, 최근접 배경화자를 찾으려면 모든 배경화자에 대해 조건확률을 계산해야 하는데 화자집단이 클 경우 계산량이 비약할 수 있고, 최대 조건 확률값으로 선택되는 최근접 화자에 따라 평준화 값이 크게 변동될 가능성이 있다.

이 문제를 보완하기 위해 화자군집 모델이 제안되었다. 이 모델에서는 주장된 화자의 인근 배경화자들(즉, 화자군집)을 다음과 같이 사용한다.

$$\log L(X) = \log p(X|S=S_c) - \log \sum_{S \in Cohort, S \neq S_c} p(X|S) \quad (3)$$

4. MLP 화자군집 모델

4.1. 기본 지역 모델

MLP는 이미 음성인식 분야에서 MLP가 가진 여러 가지 장점을 인정받아 단독[Zeppenfeld93]으로 또는 HMM과의 혼합형태[Franzini90]로 채택되고 있다. MLP의 장점을 열거하면 다음과 같다.

- 유사도 비교 방식에 비해 경쟁 집단의 거부 학습이 가능하다.
- 입력 특징의 통계적 분포에 대한 사전 지식이 필요없다.
- 고도의 병렬성과 규칙성을 가지고 있어 고성능 하드웨어 구현이 용이하다.

이 같은 장점은 화자인식에서도 그대로 적용될 수 있다[Naik94][Liou95][Fakotakis96]. 본 연구에서는 MLP의 이러한 장점과 화자군집 모델의 장점을 서로 결합시키려고 시도하였다.

분류 작업에서 MLP의 출력은 입력에 따른 출력 집단의 사후(a posteriori) 확률로 해석할 수 있다[Richard91]. 따라서 은닉 계층과 출력 계층을 각각 하나씩 갖는 MLP에서 식 (3)은 출력 계층의 노드 가운데 하나를 S_c로 지정하고 다른 노드를 S로 지정한 후 오류 역전파 알고리즘[Lippmann87]으로 학습시키는 것으로 실현할 수 있다. 그러면 학습 후 의뢰 화자의 패턴이 S_c에 속한 경우 양의 값이, *Cohort*에 속한 경우 음의 값이 출력될 것이다. 의뢰 화자의 위치가 S_c와 *Cohort*에 속하지 않는 경우 증명점수를 계산할 필요가 없다.

이러한 *Cohort*는 다음과 같은 절차를 통해 사전에 형성할 수 있다.

(1) 충분히 큰 화자집단을 학습용과 시험용으로 나눈다.
(2) 학습용을 무작위로 다시 이등분하여 A집단과 B집단으로 나눈다.
(3) A집단 내의 화자마다 MLP의 출력 노드를 할당하고 학습시킨 뒤 인식률이 일정 수준치을 넘는 것만 남기고 나머지 화자를 B집단에 넣는다.
(4) B집단의 각 화자를 학습된 MLP에 입력하여 각 출력 노드의 값을 확인한다. 이 때 출력 노드 값이 일정 수준치에 못 미치는 것을 MLP의 새로운 출력 노드에 할당하여 재학습시킨다.
(5) 새로 추가된 노드의 인식률이 (3)의 수준치 미만이면 이 화자를 다시 B집단에 넣는다.
(6) 더 이상 MLP의 출력 노드로 추가할 수 있는 화자가 없을 때까지 (4)~(5)를 반복한다.

이렇게 결정된 각 출력 노드가 개별 화자군집이 된다.

학습 과정 중에 의뢰 화자의 증명점수를 계산할 문턱값을 결정해야 한다(본 연구에서는 화자를 등록하는 학습과 구별하여 이 과정을 오프랑인 학습이라고 하고, 화자등록을 온라인 학습이라고 한다.). 이 값은 학습 화자만을 사용하여 각 화자에 대한 EER(equal error rate)을 구함으로써 결정한다. 본 연구에서는 EER을 달성하는 문턱값을 결정하기 위해 이분 탐색법(binary search)을 사용한다. 화자군집의 문턱값은 군집 내 배경화자의 문턱값을 평균하여 결정한다.

4.2. 확장 지역 모델

군집모델의 유효성은 군집 내 배경화자와의 상호관계가 충분히 반영될 수 있도록 의뢰화자의 위치가 특정 화자군집의 중심부근에 있을 것이란 가정 하에 존재한다. 그러나 실제로는 군집의 가장자리에 의뢰화자가 위치할 수 있으므로 이 경우 효과적인 모델링이 이루어지지 않으며, 그 결과 오인 거부율이 급격히 증가한다. 이 문제를 해결하기 위해 [Isobe99]는 의뢰화자의 위치에 따라 군집 모델을 재구성하는 방법을 시도하였다. 본 연구에서는 전체 화자를 대부분 대표하고 서로 상관성이 최소화된 작은 화자군집을 오프라인에서 형성하고 온라인에서 의뢰화자의 위치가 이러한 화자군집의 가장자리에 있는 경우 인근 군집들을 포함하는 더 큰 화자군집을 재구성하여 이를 학습시키는 방법을 제안한다. 이 방법의 유효성은 최초의 화자군집을 형성하는 방법에 달려있다. 즉, 서로 상관성이 적은 화자군집을 최대한 작게 설정해야 재구성 후 만들어진 화자군집의 크기가 MLP의 현실적인 온라인 학습 계산량을 만족시킬 수 있기 때문이다.

5. 실험

5.1. 데이터베이스

실험에 사용한 데이터베이스는 ETRI에서 제작한 445 PBW와 611이다. 이 둘에서 모음을 추출할 수 있는 인식기를 이용하여 남성 화자 21(445DB) + 3(611DB)명의 음성에서 /a/ 모음을 추출하였다. 16Bit, 16kHz로 샘플링되어 있는 각 음성에 32ms의 Hamming 필터를 10ms마다 씌워 프레임을 만들고 각 프레임에서 16차 Mel-scaled 필터뱅크 특징[Rabiner93]을 계산하였다. 실험화자 중 15명을 MLP 훈련에 사용하고 9명을 증명시험에 사용했다. 실험화자의 /a/ 모음 개수는 훈련화자 당 97개를 사용했고 증명화자 당 130개를 사용했다.

5.2. 시스템 구조

본 논문에서는 화자증명 문제만 다루고 있으므로 각 화자의 음성에서 음소를 인식하여 추출하는 부분을 생략한다. 또한 실험은 지속음 중에서 /a/ 모음에 한정하여 수행하였다. 시스템은 화자군집을 결정하는 MLP와 재구성된 군집에 대해 학습되는 MLP 부분으로 구성된다. 4.2절에서 설명했던 화자군집 결정 MLP(이하 MLP-1)는 미리 학습시키며 재구성 군집에 대한 MLP(이하 MLP-2)는 화자 등록시 학습시킨다.

이 실험에서 사용된 MLP는 기본적인 MLP의 구조와 오류

역전파 알고리즘과 달리 입력 계층에 특징 프레임 3개만큼의 시간지연을 두고, 한 패턴이 프레임마다 차례로 모두 입력되었을 때 MLP 내의 전체 가중치 연결이 갱신되는 방식을 취한다. 이들은 입력 계층 지연이 3이고 첫 번째 은닉 계층 지연이 1인 TDNN(time-delay neural network)[Waibel89]의 방식을 도입한 것이다. 입력 시간지연의 효과로 인식률이 향상되는 것을 [Delacretaz98]에서 확인하였으며, 위의 가중치 갱신 방식은 각 모음에 대한 시간 차이로 인해 가중치 갱신 기회에 차이가 나는 것을 막기 위해서이다.

5.3. 실험 방법 및 결과

실험은 세 방법으로 나누어 실시하였다.

(1) 기본 지역 모델
(2) 확장 지역 모델
(3) 광역 모델

광역 모델의 실험결과는 나머지 두 방법의 기준이 된다. 기본 지역 모델과 확장 지역 모델의 차이는 전자가 화자 등록시 결정된 군집에 대해서만 의뢰화자의 증명점수를 계산하는 반면 후자는 등록시 우선 소속 군집과 차선 소속 군집을 결정하여 MLP-2를 형성할 때 이들을 합하여 화자군집을 재구성한다는 것이다.

지역 모델의 MLP-1은 20개의 은닉 노드와 3개의 출력 노드를 갖고 있으므로 3개의 화자군집을 모델링한다. MLP-2는 40개의 은닉 노드와 1개의 출력 노드를 갖고 있다. 광역 모델의 MLP는 40개의 은닉 노드와 1개의 출력 노드를 갖는다. 각 은닉 노드의 수는 반복 실험을 통해 결정했다. MLP-1의 각 화자군집에는 5명의 배경화자가 할당되고, 광역 모델의 MLP에도 이들이 할당되어 등록시 의뢰화자의 경쟁 학습에 사용된다. 학습시 97개 패턴 중 10개를 사용하고 나머지 87개는 MLP 파라미터와 문턱값 결정에 사용했다. 증명시 의뢰화자의 130개 패턴 중 10개를 학습에 사용하고 120개 패턴과 나머지 시험화자의 모든 패턴을 오류율 측정에 사용했다.

MLP의 분류 결정 기준은 MLP-1의 경우 우세한 군집의 특징 프레임 수이고, MLP-2의 경우 50%를 기준으로 문턱값을 넘는 특징 프레임 수이다. 시험 화자 9명에 대한 오류률은 오인 거부율(false rejection)과 오인 수락률(false acceptance)로 나누고 /a/ 모음 개수를 1에서 5개까지 늘려가며 측정하였고 그 결과를 표1에 정리하였다.

/a/ 수	기본지역모델		확장지역모델		광역모델	
	FR	FA	FR	FA	FR	FA
1	40.6%	11.4%	14.3%	14.1%	10.8%	12.8%
2	35.6%	11.9%	9.8%	14.0%	7.3%	11.2%
3	32.2%	11.7%	7.1%	13.9%	5.1%	10.6%
4	28.0%	11.3%	6.1%	13.9%	5.0%	10.3%
5	29.4%	10.4%	4.4%	13.8%	4.8%	10.3%

표1. 시험 화자에 대한 오류율 결과

결과에서 볼 수 있듯이 광역 모델을 기준으로 기본 지역 모델의 오인 거부율이 크게 떨어졌다. 확장 지역 모델에서는 기본 지역 모델에 비해 오인 거부율이 개선되는 대신 오인 수락률이 약간 증가하는 것을 볼 수 있다. 전반적으로 오인 수락률이 높은 것은 사용한 데이터베이스의 화자수가 적기 때문인 것으로 보인다. 그러나 이 결과에서 기본 지역 모델에 비한 확장 지역 모델의 효과를 확인할 수 있다. 그리고 광역 모델에 비해 기본 지역 모델은 1/3, 확장 지역 모델은 2/3의 학습 화자만 사용함으로써 학습 계산량을 감소시킬 수 있었다.

6. 결론

본 논문에서는 기존의 확률적 화자군집 모델을 MLP로 구현하는 방법과 원형 화자군집 모델이 갖는 문제를 해결할 수정 모델을 제시했다. 화자군집 모델은 화자등록 시간에 민감한 실용 환경에서 중요한 의미를 지닌다. 이 외에도 본 연구에서는 지속음을 인식단위로 하여 어구독립 화자증명을 지향하였으며 차후 부단어 단위 음성인식 기능을 추가함으로써 어구요구 방식을 달성할 수 있도록 하였다. 비록 실험에서는 /a/ 모음에 대한 결과만 제시했지만 화자증명에 사용하는 평균적인 어구에서 지속음을 추출할 비음, 모음, 마찰음의 수가 실험에서보다 더 많을 것이므로 더 낮은 오류율을 기대할 수 있다.

7. 참고 문헌

[Delacretaz98] D. P. Delacretaz and J. Hennebert, "Text-Prompted Speaker Verification Experiments with Phoneme Specific MLPs," ICASSP, Vol. 2, pp. 777~780, 1998.

[Eatock94] J. P. Eatock and J. S. Mason, "A Quantitative Assessment of the Relative Speaker Discriminating Properties of Phonemes," ICASSP, Vol. 1, pp. 133~136, 1994.

[Fakotakis96] N. Fakotakis and J. Sirigos, "A High Performance Text Independent Speaker Recognition System Based on Vowel Spotting and Neural Nets," ICASSP, Vol. 2, pp. 661~664, 1996.

[Franzini90] M. Franzini, et al., "Connectionist Viterbi Training: A New Hybrid Method for Continuous Speech Recognition," ICASSP, Vol. 1, pp. 425~428, 1990.

[Furui96] S. Furui, "An Overview of Speaker Recognition Technology," Automatic Speech and Speaker Recognition, Kluwer Academic Publishers, pp. 31~56, 1996.

[Gu99] Y. Gu and T. Thomas, "A Hybrid Score Measurement for HMM-Based Speaker Verification," ICASSP, Vol. 1, 317~320, 1999.

[Isobe99] T. Isobe and J. Takahashi, "A New Cohort Normalization Using Local Acoustic Information for Speaker Verification," ICASSP, Vol. 2, pp. 841~844, 1999.

[Liou95] H. Liou and R. J. Mammone, "Speaker Verification Using Phoneme-Based Neural Tree Networks and Phonetic Weighting Scoring Method," Proceedings of the 1995 IEEE Workshop Neural Networks for Signal Processing V, pp. 213~222, 1995.

[Lippmann87] R. P. Lippmann, "An Introduction to Computing with Neural Nets," IEEE ASSP Magazine, Apr 1987.

[Naik94] J. M. Naik and D. M. Lubensky, "A Hybrid HMM-MLP Speaker Verification Algorithm for Telephone Speech," ICASSP, Vol. 1, pp. 153~156, 1994.

[Rabiner93] L. Rabiner and B. Juang, Fundamentals of Speech Recognition, Prentice Hall, 1993.

[Richard91] M. D. Richard and R. P. Lippmann, "Neural Network Classifiers Estimate Bayesian a posteriori Probabilities," Neural Computation, No. 3, pp. 461~483, 1991.

[Waibel89] A. Waibel, et al., "Phoneme Recognition Using Time-Delay Neural Networks," ASSP, Vol. 37, pp. 328~339, Mar 1989.

[Zeppenfel93] T. Zeppenfeld, et al., "Improving the MS-TDNN for Word Spotting," ICASSP, Vol. 2, pp. 475~478, 1993.